DOCUMENT RESUME

ED 178 583 TM 009 753

AUTHOR	Rudner, Lawrence M.; And Others
TITLE	The Problem of Item Bias: A Comparison of Techniques. Final Report.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
PUB DATE	[79]
GRANT	NIE-G-78-0084
NOTE	62p.
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Analysis of Variance; *Comparative Statistics; Complexity Level; Factor Analysis; *Item Analysis; Simulation; *Test Bias; Test Items; Test Results
IDENTIFIERS	Chi Square Test; Item Characteristic Curve Theory; Item Discrimination (Tests); Monte Carlo Method

ABSTRACT

The psychometric rationales of six techniques for quantifying estimated item bias were reviewed: analysis of variance, transformed item difficulties, chi square, item characteristic curve theory, factor analysis, and distractor response analysis. The limitations and advantages of the approaches were described in terms of their underlying assumptions and psychometric soundness. By using Monte Carlo generated item response data, theis research also sought to determine the effectiveness, sufficiency and similarity of select biased item detection techniques. Experimental variables included: different proportions of bias for items in the item pool, different types of generated item bias, and different test lengths. The three parameter item characteristic curve (ICC) theory and the five interval chi-square techniques were found to produce satisfactory results under all test conditions. The transformed item difficulties technique was found to be sensitive to bias in item difficulty but insensitive to bias in item discrimination. The one parameter ICC theory technique was unsatisfactory under all conditions. (Computer program subroutines and bibliographic references are appended). (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED178583

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

The Problem of Item Bias:

A Comparison of Techniques

Final Report

Lawrence M. Rudner

Pamela R. Getson

David L. Knight

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

LAWRENCE M. RUDNER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

TM009253

The authors gratefully acknowledge the helpful comments and editorial assistance of William R. Merz in preparing this manuscript.

This research was sponsored in part with funds from the National Institute of Education Grant NIE-G-78-0084 under Project 8-0460. The views and opinions expressed herein are those of the authors' and do not necessarily reflect those of the National Institute of Education, Department of Health, Education and Welfare.

Abstract

The use of certain diagnostic and achievement tests have recently been viewed as contradictory to a basic civil right - the fair treatment of individuals regardless of race, sex, religion or national origin. By identifying and removing biased items in test development, this right can be fostered.

This paper reviews the psychometric rationales of the six classes of techniques for quantifying item bias: (1) analysis of variance approaches in which bias is operationally defined in terms of significant item by group interaction effects; (2) transformed item difficulties approaches in which within group p-values are standardized and compared between groups; (3) chi-square approaches in which individual items are investigated in terms of between group score level differences in expected and observed proportions of correct responses; (4) item characteristic curve theory approaches in which differences in the probabilities of a correct response, given examinees of the same underlying ability and different culture groups, are evaluated; (5) factor analytic approaches in which item bias is investigated in terms of culture specific and culture common sources of variance or in terms of loadings on a biased test factor; and (6) distractor response analysis approaches in which the relative attractiveness of item distractors are investigated. Limitations and advantages of the approaches in terms of their underlying assumptions and psychometric soundness are discussed.

By using Monte Carlo generated item response data, this research also sought to determine the effectiveness, sufficiency and similarity of select biased item detection techniques. Experimental variables included: different proportions of bias for items in the item pool, different types of generated item bias, and different test lengths. The three parameter ICC theory and the five interval chi-square techniques were found to produce satisfactory results under all test conditions. The transformed item difficulties technique was found to be sensitive to bias in item difficulty but insensitive to bias in item discrimination. The one parameter ICC theory technique was unsatisfactory under all conditions.

TABLE OF CONTENTS

Abstract ii

I. INTRODUCTION 1

Background 1

Concept of Item Bias 2

II. APPROACHES TO BIASED ITEM DETECTION 4

Analysis of Variance 4

Transformed Item Difficulties 6

Chi-square 11

Item Characteristic Curve Theory 12

Factor Analytic 15

Distractor Response Analysis 17

Discussion of Approaches 20

III. METHOD 23

Data Generation 23

Application of the Approaches 27

IV. RESULTS 30

V. DISCUSSION 36

VI. CONCLUSION 39

Appendix - FORTRAN program for data generation and analysis . . 40

Authors 52

References 53

CHAPTER 1

INTRODUCTION

Background

Approximately 25 years ago Eells and his colleagues conducted what appears to be the first serious attempt to examine test items for bias (Eells, et. al., 1951) and developed one of the first measures purported to be culture fair. Since that time, the entire issue of cultural bias in measurement has become heated, complex, and pronounced in the literature. Actions by the National Association of Black Psychologists, the American Personnel and Guidance Association, the National Education Association, the National Association for the advancement of Colored People, the National Association of Elementary School Principals, and the Council of the Society for the Psychological Study of Social Issues call for moritoria on certain types of tests, ban other tests, and by requiring alternate plans for testing, indicate the serious nature of the current situation (see Williams, Mosby and Hinsen, 1978). The concern is also apparent in recent litigation (Defunis vs. Odegaard, 1974; Diana vs. the California State Board of Education, 1970; Hobsen vs. Hansen, 1967). Naturally, all this has not gone unnoticed by those involved in the measurement field. Bias and debiasing studies have been conducted and various models have been proposed in expanding efforts to meet the challenge of reducing or minimizing bias in educational assessment.

One major type of bias investigation is concerned with the fairness of the instrument as a whole and examines the question: "Does a test either unduly favor or impede examinees from different parts of the

country or from different backgrounds?" Another type of investigation is concerned with the items within a test and asks: "Which items and item formats are appropriate for a given population and which may be used across all specified groups?"

The first type of investigation is of interest to test users who need to know the accuracy of specific test information. The models proposed by Cleary (1968), Thorndike (1971), Darlington (1971), Cole (1973), Einhorn and Bass (1971) and Gross and Su (1975) exemplify the first type of investigation and are referenced as concerning test bias.

The second type of investigation is of interest to test developers as it assists them in developing valid and cross-culture fair items while providing a framework for constructing better tests in subsequent efforts. The work of Angoff (1972), Cardall and Coffman (1964), Green and Draper (1972), Merz (1973, 1976a), Rudner (1977a), Scheuneman (1975) and Veale and Foreman (1975, 1976) are examples of this type of investigation . It is this second type of investigation, referenced as concerning item bias, which is discussed in this paper.

Concept of Item Bias

In theory, tests and test items are perfectly unidimensional, i.e., an item measures only one ability and all items of a test measure the same ability. In practice, however, this absolute is never attained. A mathematics item, for example, may rely for its solution upon a specific amount of reading competence as well as knowledge of mathetmatical operations. When the item is administered to a group whose reading ability is greater than that required by the item, the reading component becomes irrelevant. The issue of item bias enters into consideration when such an item is administered to two different groups of examinees, and the

responses of one of the groups are dependent on the secondary skill. An item designed to assess mathematics ability and, in practice, behaves as a mathematics item for one group of examinees and as a combination of mathematics and reading item for another group of examinees is considered biased. Item bias, then, refers to the extent an item measures a set of skills differently across different populations.

The issue of item bias is distinct from the issue of test bias, even though item bias may contribute to test bias because a biased item may favor one population group. However, since a test is composed of many items, it is entirely possible for the amount of item bias favoring one group to equal the amount of item bias favoring another group. Hence, the bias in the items may not affect aggregate group performance. Bias in test items is also not necessarily directional. An item which favors high ability examinees in one group and low ability examinees in another would be considered biased. The net effect of such an item, however, may not be evident in overall item difficulty. Additionally, item bias does not address how a test is used.

Practitioners often desire to dichotomously classify items as either biased or unbiased. Item bias by definition, however, is a matter of degree and any reference to an item being biased or unbiased simply refers to whether the bias in an item is considered to be relatively minimal or substantial.

CHAPTER II

APPROACHES TO BIASED ITEM IDENTIFICATION

Recently, six major categories of procedures and models have been proposed and advocated for identifying biased items within a test: (1) (1) analysis of variance approaches, (2) transformed item difficulties (p-values) approaches, (3) chi-square approaches, (4) item characteristic curve theory approaches, (5) factor analytic approaches, and (6) distractor response analysis approaches.

Analysis of Variance Approaches

In this first type of approach, individuals sampled from two or more populations are given a common test and the resultant variations in item scores are analyzed by an analysis of variance design. Item score variance could be attributed to differences in (1) items, as some items are more difficult than others; (2) groups, as one group may have more ability in the measured attribute than another; (3) subjects within groups, as examinees will differ in ability; and (4) interaction of the items and the groups. In this technique, bias is defined as a significant item by group interaction, i.e., the presence of items which are relatively more difficult for members of one group than another. Examples of this approach can be found in Cardall and Coffman (1965), Cleary and Hilton (1968), Hoepfner and Strickland (1972), and Jensen (1973).

Application of this procedure to item test data can be accomplished in the standard manner providing that certain assumptions are met. In order to meet the homogeneity of variance assumption, Cardall and Coffman

(1964) transformed the within group item difficulties with an arcsin transformation. Plake and Hoover (1977) extended this technique to allow for the identification of individual biased items. The interaction contrast for each item within each group then takes the form:

$$\Psi_{ij} = \theta_{ij} - \theta_{i.} - \theta_{.j} + \theta_{..}$$

where θ_{ij} is the arcsin transformed item difficulty for the i^{th} and the j^{th} group.

The error variance is:

$$\sigma\Psi_{ij} = \frac{q-1}{2q\,\tilde{n}_j}$$

where q is the number of items

$\tilde{n}_j$ is the harmonic mean of the number of subjects in the j^{th} group simultaneous significance test, such as Bonferoni's procedure, can then be used to identify individual items which appear to contribute differentially to the variance and are, therefore, considered biased.

While this technique can readily be applied to any data, Hunter (1975) illustrated a critical problem in showing that item by group interactions will occur in completely unbiased tests merely as a function of differing levels of ability or mean performance. The item means for two groups of unequal ability will differ by a different amount depending on the overall difficulty of the item. Medium overall difficulty items will show larger differences than harder or simpler ones. Items of different overall difficulty, therefore, will always "show an 'item by group' interaction in any situation in which the two groups differ in achievement level." (p. 10)

Jensen (1973) reported studies in which he attempted to control for different mean levels of performance by matching subjects on their mental

age. In the first study, great reductions were found in the item by culture interaction after matching, indicating that the procedure may be more sensitivie to ability than to cultural variations. Jensen confirmed these results in a second investigation. After using an analysis of variance approach with white and black subjects (without matching mental age), he conducted a second analysis of variance using two groups of Caucasians whose score distributions closely matched those of the blacks and whites in the first part of the study. The results of this pseudo-race comparison closely matched those of the true race comparison, especially with regard to the item by culture interaction. He concluded that "it would be extremely difficult to make a case that the race by items interaction is attributable to cultural bias" (p. 17).

When a significant interaction is found there is the additional and practical problem of choosing an appropriate post-hoc test procedure. If the procedure selected does not control for family alpha level, the actual alpha level will become greatly inflated as the number of items increases. On the other hand, if a post-hoc procedure is employed which does control for the family alpha level (e.g., the Bonferoni statistic) the reduction in true alpha level would require relatively large differences for statistical significance. Consequently, only extremely biased items would be identified, and the sensitivity of the procedure would create a questionable degree of utility.

Transformed Item Difficulties

The transformed item difficulties approach, providing for a visual examination of item by group interaction effects, was probably first described by Thurstone (1925) in connection with his method of absolute scaling. Of all the approaches, this method appears to be one of the

most widely known. It has been used by Angoff (1972; Angoff and Ford, 1974; Angoff and Modu, 1973) and by others (Green and Draper, 1972; Jensen, 1973; Hicks, et al. 1976; Strassberg-Rosenberg and Donlon, 1975; Echternacht, 1974; and Rudner, 1978). Further, the approach has appeared in at least one measurement textbook (Anastasi, 1976. pp. 222-226).

Angoff (1972) states that:

> "In this method, indices of item difficulty; i.e., p-values are obtained for two different groups on a number of items. Each p-value is converted to a normal deviate and the pairs of normal deviates, one pair for each item, are plotted on a bivariate graph, each pair represented by a point on the graph. (p. 1).

The plot will generally form an ellipse. A 45 degree line, passed through the origin, describes the theoretical relationship which indicates the absence of bias. Items deviating greatly from this line may be regarded as exhibiting an item by group interaction. Relative to the other items, deviant items are especially more difficult for members of one group than the other. Assuming both groups received similar instructions, such items would appear to represent the measurement of different concepts for the two groups of examinees.

The distance of an item point to the line, (d) serves to indicate the degree of item bias.

$$d = (z_1 - z_2)/\sqrt{2}$$

where z_1 and z_2 are the transformed item difficulty for groups 1 and 2 respectively

Items which are "greatly deviating" from the line are identified by a method of outlier or residual analysis. One method places confidence limits on the line by using a multiple of the standard error of

estimation. An alternate approach, adopted by Strassberg-Rosenberg and Donlon (1975) and Hicks, et al. (1976) involves computing the standard deviation of the residuals and classifying as biased those items deviating by greater than 1.5 standard deviations. Rudner (1978) has employed a fixed item-regression line distance of .75 z-score units.

Since the intent of transformed item difficulty approaches is to make comparisons of between-group differences in item difficulty, it is necessary to transform the proportion passing an item to an index of item difficulty which constitutes at least an interval scale. This is accomplished by expressing each item p-value in terms of within-group deviations of a normal curve (cf. Guilford, 1954, pp. 418-419). Any linear transformation of the item z-score will meet such a requirement. Delta values (4z + 13) represent one such transformation which has been applied frequently.

An example of the fixed item-line regression approach utilizing score transformations is shown in Figure 2.1 The transformed p-values have a correlation of approximately .90, making the plot relatively long and flat. The solid line represents the main axis and the dotted lines represent linear confidence limits. The item represented in the upper left quadrant, outside the confidence interval, would be considered biased.

As a modification of this procedure, Green and Draper (1972, p. 16) suggest that the "item-test biserial correlations might be incorporated . . . so as to estimate the linear test score-item score regression whereby item difficulties may be formed in a manner analogous to the way in which adjusted means are formed in an analysis of covariance." Since by this procedure, differential item discrimination indices and item

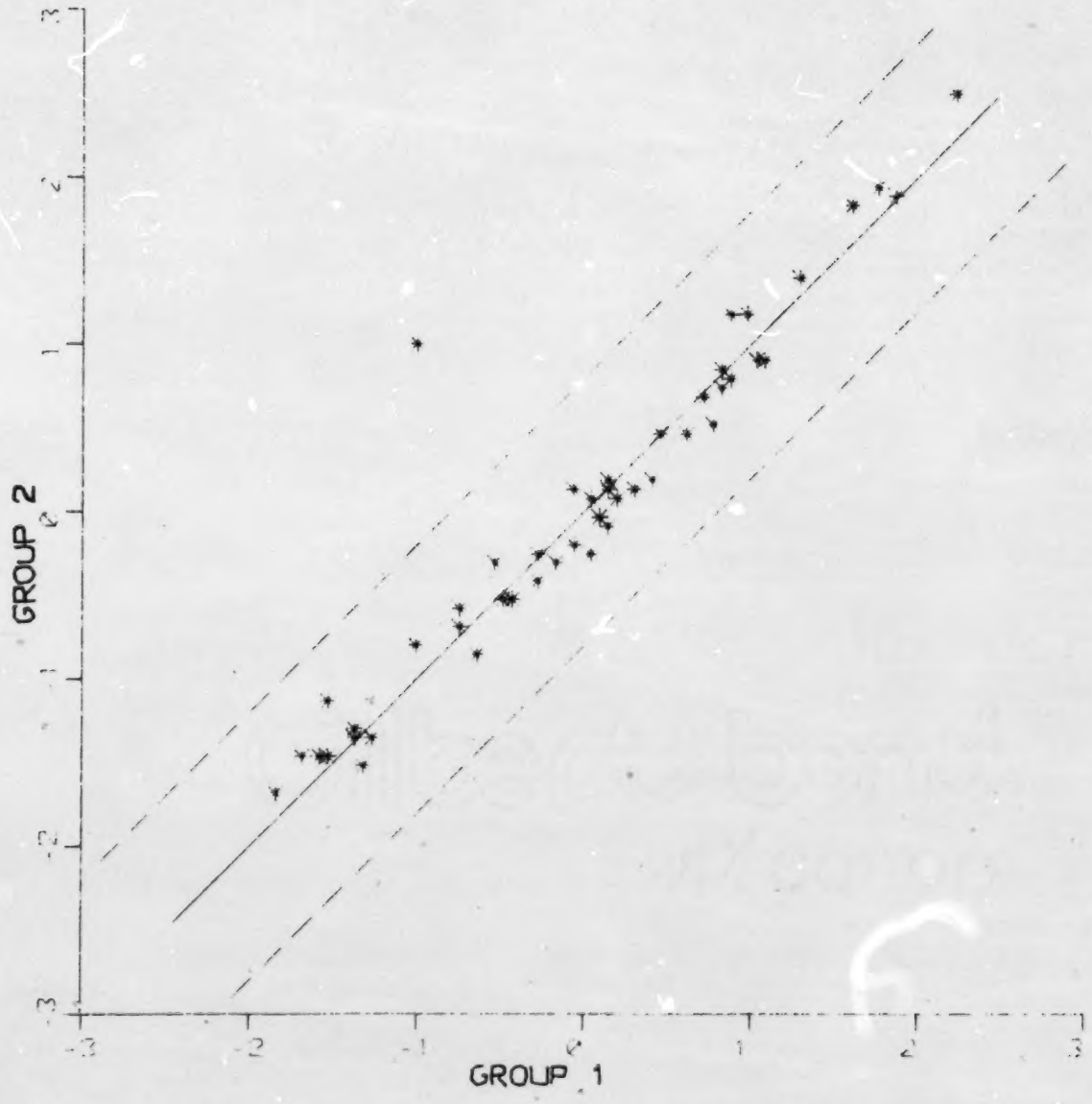


Figure 2.1: A hypothetical transformed item difficulties scatterplot

difficulties would both influence item locations on the regression plot, items which have proportional p-values but disproportionate discrimination indices would have a greater tendency to deviate from the main axis of the scatterplot and show up as aberrant.

Another modification of the transformed item difficulty approach was that presented by Fishbein (1974). The Fishbein technique calls for the calculation of a mean deviation p-value for the groups to be compared. This average deviation is then compared with the individual item p-value. The difference (either positive or negative) serves as an index of bias. That is, those items with the greatest positive or negative deviation from the mean deviation are considered most biased.

Echternacht's (1974) modification to the fixed item-line regression approach also begins with item difficulties transformed to delta values. Differences in transformed item difficulties are then computed for each pair of groups and plotted on normal probability paper. Additionally, a line is plotted to represent a hypothetical normal distribution with the obtained mean and standard deviation of the difference between pairs as parameters. Confidence bands constructed around this hypothetical normal distribution line represent the area outside of which biased items would fall.

It should be noted that the transformed item difficulty and the analysis of variance approaches are distinguished primarily by the way in which they quantify the item by group interaction. Both approaches define bias as an item by group interaction; the transformed item difficulty approaches identify items as biased relative to each other, analysis of variance approaches quantify bias through overall significance testing. The problems associated with an item by group

interaction compounded with differential ability noted by Hunter (1975) and reported in the analysis of variance section, apply to transformed item difficulty approaches as well.

Transformed item difficulty techniques are based on p-values which index the average difficulty of an item for a given group. As a result, only items which are illustrative of differential aggregate group performance will be identified. While this is probably the type of biased item of greatest social concern, it is not the only kind of biased item.

Chi-square Approaches

A third approach to biased item analysis determines whether examinees of the same ability level have the same probability of a correct response regardless of group affiliation. Each tryout sample is blocked into a specified number of groups based on observed scores the proportions of students within each level responding correctly are compared. An item is considered unbiased if, for all individuals in the same total score interval, the proportion of correct responses is the same for both groups under consideration.

Scheuneman (1979) advocates a modified chi-square to estimate the probability that the item is unbiased by the definition above. The expected values for each cell (E_{ij}) are obtained by multiplying (1) the proportion of all examinees with total scores within interval j responding correctly to the item by (2) the number of examinees within the cell. That is,

$$E_{ij} = \frac{O_{.j}}{N_{.j}} N_{ij}$$

where $O_{.j}$ is the number of examinees in total score interval j responding correctly
$N_{.j}$ total number of examinees in total score interval j
N_{ij} is the total number of examinees in group i and score interval j

As in a conventional chi-square, observed cell values are simply the number of examinees within the cell responding correctly to the item.

Rudner and Convey (1978) show that the total score distribution of the two groups affect the expected cell values. As a result, the X^2 values will be inflated when the observed score distributions are different. The inflation of the X^2 values will be systematic when identical intervals are used for each item. Systematic inflation allows the magnitude of the X^2 to be used as a relative index of bias. The probabilities associated with the X^2, however, will be misleading.

The chi-square approach has the advantage of examining items for bias based on the distributions of correct responses across ability levels. As a result, the technique is not limited to identifying bias only when it favors one entire group.

Item Characteristic Curve Theory Approaches

Item characteristic curve (icc) theory relates the probability of a correct item response to a function of an examinee's underlying ability level (θ_j) and the characteristic(s) of the item. While the various models (Lord, 1952: Rasch, 1960; Birnbaum, 1968; Urry, 1970) differ in terms of the number of item parameters considered, they all describe the item parameter(s) independently of the examined sample. This attractive property has led to the development of some interesting applications in test development, adaptive testing and equating, and may prove useful in detecting item bias.

In an earlier study, Green and Draper (1972) used observed total scores as estimates of examinees' abilities, and the proportions of examinees responding correctly at each total score level as estimates of the probability of a correct response to item g given θ_j, P(ug=1). Their procedure called for plotting estimated icc's for each item separately

for each culture group and then comparing the plots.

By this and other icc theory approaches, an item is unbiased if examinees of the same ability level, but of different group affiliations, have equal probabilities of responding correctly. That is, an item is unbiased if the estimated icc's obtained from the various groups are identical. As an example of a biased item, consider the two hypothetical curves shown in Figure 2.2. These curves are based on responses made by two different groups of examinees to the same item. Total observed scores are placed on the x-axis and the proportions of examinees in groups 1 and 2 responding correctly to the item are placed on the y-axis. For this item, the proportion of examinees responding correctly is a function of group membership in addition to being a function of ability as estimated by total observed score. Such an item is considered biased in that often examinees of the same ability level, but from different groups, do not have similar proportions of correct responses.

While this approach by Green and Draper has appealing features, total observed scores are directly incorporated and quantification of the degree of item bias is difficult since subjective visual observation is used to identify a "very biased item." Additionally, as Hunter (1975) has correctly indicated, the interaction used to identify biased items stems from the nonlinearity of the observed icc's which is a function of the observed total score distribution.

Rudner (1977a), Lord (1977), and Pine (1976) have refined the procedure used by Green and Draper to identify biased items by incorporating equated icc parameter values for the three parameter Birnbaum (1968) model. Rudner used the area between pairs of equated icc's to indicate the relative amount of aberrance of each item, as well as visual

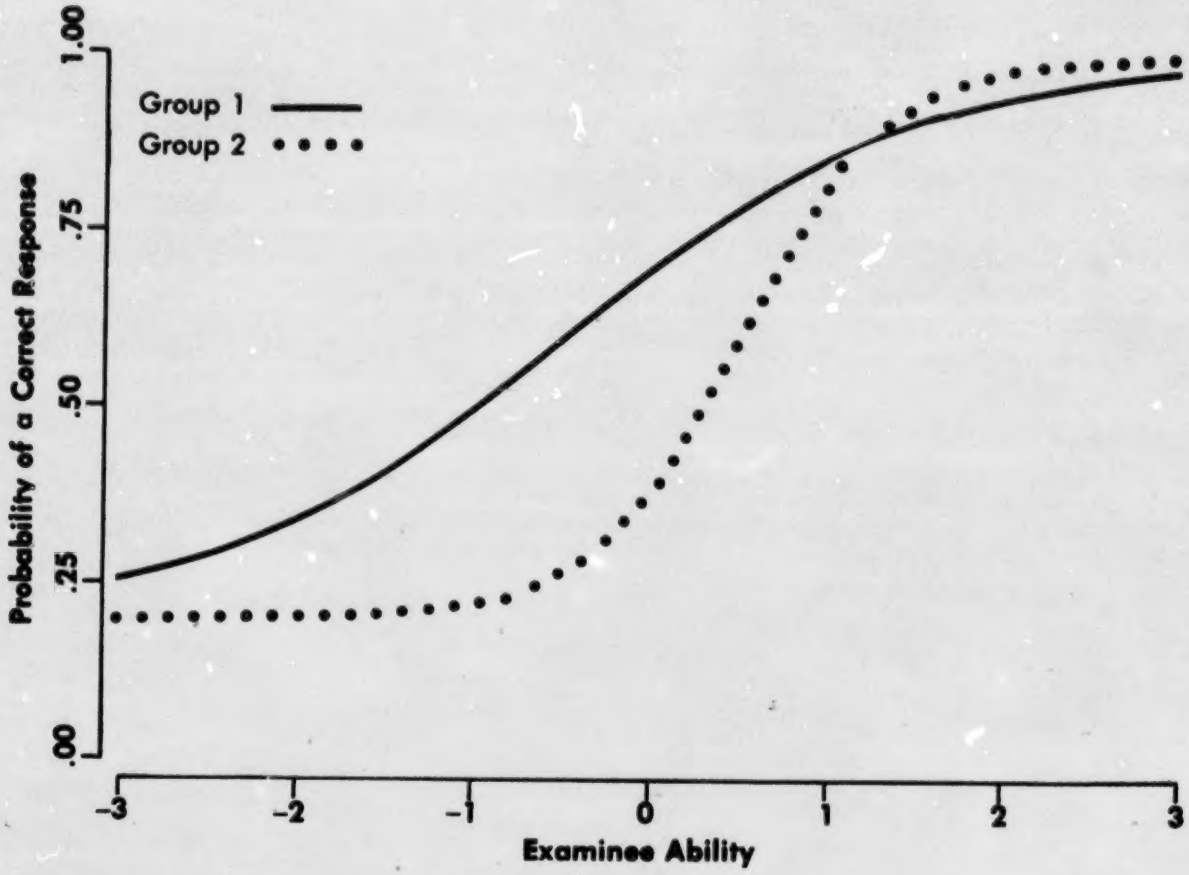


Figure 2.2: A hypothetical biased test item

inspection of the equated icc's to provide additional information pertaining to the nature of the aberrance. Lord has employed an asymptotic significance test based on the summed variance-covariance matrices of the equated parameter estimates to test for significant differences between pairs of equated icc's. Pine used the residuals from equating the difficulty parameters as an index of item aberrance.

Using the one parameter Rasch model, Durovic (1976) and Wright, Mead and Draba (1976) focus their attention on differences in the relative difficulty of the items. The differences between the observed item responses and the predicted probabilities of a correct response are computed. The goodness-of-fit residual is then analyzed for between-group differences. Like the chi-square technique, the three parameter icc theory techniques are appealing in that they are not limited to identifying bias only when it favors one group. As exemplified in Figure 2, an item which favors high ability examinees in one group and low ability examinees in the other can be identified.

The major limitation of the icc theory techniques is that they assume perfectly accurate estimates of the probability of a correct response to each item. In practice, however, parameters can not be determined for all items; therefore not all items can be evaluated. Additionally, the parameter estimates are never free from error and consequently, dissimilar equated icc's may be indicative of sampling fluctuations as opposed to item bias.

Factor Analytic Approaches

In factor analysis, underlying factors (i.e., dimensions or traits) are hypothesized and correlations of each variable with the hypothesized factors are computed. In applying factor analysis to an achievement test

each item is treated as a variable. Such an analysis could be conducted twice using examinees from two different cultural backgrounds. Ideally, the two separate groups of examinees would yield similar sets of item-trait correlations (factor loadings). Different sets of factor loadings would indicate that the two groups are not responding to the items in the same manner. Such a test would be considered biased in that it appears to measure a different trait across groups. The items exhibiting the most bias would then be those with the largest differences in factor loading.

Green and Draper (1972) and Green (1976) suggest an inter-group factor analysis model based on the inter-battery factor analysis approach offered by Tucker (1958). In this inter-group model, the item variance is partitioned into: (1) factors common to each subgroup; (2) factors specific to subgroups; and (3) residual or error variance. With this model, one can determine the proportion of item variance accounted for by a given subgroup. An item, then, is unbiased when this proportion is small and biased if a large proportion of variance is attributable to group-specific courses.

Merz (1973, 1976a) developed an alternate approach which incorporates factor scores and analysis of variance. The item intercorrelation matrix is computed for subjects pooled across groups. The matrix is reduced with principal components analysis, employing an eigenvalue criteria of 1.0 to determine the number of factors to be extracted. The factor matrix is then rotated orthogonally to simple structure, and factor scores derived from the rotated matrix. An analysis of variance is then conducted on each set of factor scores using group membership as the independent variable and factor scores for each

vector as the dependent variable. Item bias is defined as a major loading on a factor with a significant F ratio on a main effect or on an interaction.

In general, these factor analytic approaches are appealing in that they deal with the underlying true abilities of the examinees. Green and Draper's approach (1972) is particularly appealing in that variance is partitioned into culture-specific and culture-common sources. Merz's approach (1973) has an advantage in that variance attributable to factors such as socio-economic status, IQ, and sex can be partialed out.

The decision problems that beset factor analysis in general are multiplied when applying these factor analytic techniques to test bias. The user must decide on the type of correlation matrix to use, the type of factor analysis, the number of factors to extract, and the type of rotational scheme. Different sets of decisions could lead to different results. In applying the factor score approach, Rudner and Convey (1978) found the procedures they used to yield unsatisfactory results. A large number of items were found to be biased in a pseudoculture group comparison, and the variance of the detected amount of bias in an actual application was minimal.

Distractor Response Analysis

Veale and Foreman (1975, 1976) recommend investigating the distractor response distribution for various cultural groups in an approach not dependent on total test scores. Should one group be overly attracted to a particular distractor in comparison to a second group, there may be a biasing characteristic of the item attracting them away from the correct response. For this approach, bias is defined as those characteristics of an item which cause a distortion in the item p-value

for a specific group.

Consider the choice distributions illustrated in Table 1. Different proportions of examinees responded correctly to the item and different proportions of examinees selected each of the three distractors of incorrect alternatives. For each item, Veale and Foreman test whether there is an association between distractor selected and culture group. A chi-square is computed from a two-way table of groups by distractors with cell entries being the "pull index" of the distractor, i.e., the percent of examinees in the group responding incorrectly and selecting the distractor.

Insert Table 1 about here

Maw (1979) has developed an alternate approach based on the work of Ku and Kullback (1974). For each item, a contengency table is developed which includes a breakdown of item distractors by groups as is done by Veale and Foreman. Maw, however, includes additional information on variable which are known correlates of educational achievement; e.g., home background and attitudes, instructional processes, and socio-economic status. These known correlates expected are to account for a large proportion of the item variance. Various loglinear models are fit to the data until the data are adequately represented. The resultant parameters of the model are then investigated for information relating to distractor patterns. Biased items are identified by a significant distractor-by-group marginal effect. The individual parameters of the marginal effects are then analyzed to determine the contributions of the various item response choices.

Analysis of distractor bias is appealing because it avoids the

Table 1

Hypothetical Item Alternative Distributions for

Two Groups of Examinees

Percent of Examinees in Two Groups Selecting Each Alternative to a

Hypothetical Multiple Choice Item

	Correct Response %	Distractor 1 % pull	2 % pull	3 % pull
Group 1	45	6(.12)	14(.28)	30(.60)
Group 2	73	4(.08)	10(.45)	8(.36)
Total	68	5	11	1s .

Adapted with permission from Veale and Foreman, 1976

assumption inherent in other techniques that total test scores are valid indicators of ability. The technique can help identify particular distractors causing bias and, therefore, offer an alternative to the practice of discarding intact items in order to debias a measure. Finally, distractor response analysis can be applied to criterion referenced tests which have an expected restriction in range.

Since these distractor response analysis techniques are dependent on inference testing, even small differences between observed and expected cell frequencies can accumulate to achieve statistical significance. Consequently, a relatively unbiased item may appear aberrant. Veale's technique (1975) is only capable of identifying bias which is manifested in the incorrect responses. Maw's technique requires the collection of sizeable amounts of data.

Discussion of Approaches

Several approaches for identifying biased items have been presented along with the rationals for their developmental and application and their more apparent advantages and disadvantages. Any attempt to apply one of these approaches in a real life situation must consider factors besides the psychometric characteristics of the technique. In practice, the best theoretical technique may become the worst and positive procedural benefits may turn into limitations. A specific model may be impossible to apply or statistical requirements, may not be met.

Regardless of the specific application, item bias detection techniques are designed to identify items which are not uniformly answered by examinees in different groups. This does not mean that items identified as biased should, by definition, be considered bad and removed from a measure. In some situations biased items will be the most informative

and the best predictors, consequently making them of great interest. For example, the items identified as most biased may be used for focusing educational priorities for certain groups of students.

The degree of item bias, variously indicated by the magnitude of a residual, area, distance, factor loading, χ^2 or F ratio, should be considered along with indices of item difficulty, discrimination, and content when an item is evaluated. The severity of the consequences of misclassification as well as the distribution of bias within and across measures must aid in determining the classification of an item as biased or unbiased. Some techniques offer significance testing as a classification scheme and appear to relieve the practitioner of responsibility. However, such significance testing ignores the fact that bias is a matter of degree.

Regardless of the purpose or method of applying an item bias detection techniques, the practitioner must recognize that all of the item bias detection techniques have inherent problems and limitations. First, all of the techniques appear to be sensitive to factors such as differences in group ability distributions, imperfect parameter estimates and violations of unidimensionality, which are extraneous to the problem of item bias. Some of the techniques appear to be insensitive to certain types of bias, such as bias in item discriminatory ability. In addition, the practitioner must concurrently consider the prohibitive costs of applying some of the techniques which require vast numbers of examinees, items, demographic data, and large amounts of expensive computer and personnel time. Finally, certain methods are extremely complex and cannot be explained to the average consumer.

While problems do exist with the methodologies the problems of bias in test items exist in any situation where a measure is given to examinees from more than one culture or population group. Under this circumstance, then, all of the techniques reviewed in this paper have specific strengths to encourage their use in investigating item bias. Some methodologies will be more appealing on theoretical grounds, others on practical grounds. It must be recognized that the specific items identified as most biased will vary depending on the technique chosen, since each technique differs in both the manner in which item bias is defined and quantified. In addition, indiscriminant removal of items identified as most biased will not necessarily insure fair or appropriate use of measures and may not substantially alter group mean scores.

In summary, item bias detection techniques seek to provide for the quantification of the degree to which an item measures a set of skills differently across identifiable population groups. If needed, this quantification of bias permits the user to set cut-points for 9 dichotomous classification of items as biased or unbiased. While none of the methods is capable of identifying all instances of biased items and all have their problems and limitations, they can provide important additional information for test users and developers concerned with this sensitive measurement issue.

CHAPTER III

METHODOLOGY

Data Generation

The generation of the data for this study consisted of two parts

1) the specification and generation of biased test items,

2) the specification and generation of examinees and their responses to the biased items.

Birnbaum's (1968) three parameter cumulative logistic model was chosen to relate item and examinee characteristics to item responses.

Birnbaum's model states that the probability of a correct response to a single item g, $P(u_g=1)$, is a function of three item parameters - item discrimination (a_g), item difficulty (b_g), pseudo-guessing (c_g), and one examinee parameter - ability (θ_i).

$$P(u_g=1 \mid \theta_i) = c_g+(1-c_g)[1+\exp(-1.7a_g(\theta_i-b_g))]^{-1} \qquad (1)$$

where: i indexes examinees, and
g indexes items

Once the parameters for the item are specified, the probability of a correct response is solely a function of examinee ability. As examinee ability increases, so does the probability of a correct response. With this model, the probability of a correct response for an examinee is determined once the ability for the examinee is defined. The probability of a correct response can be transformed to an observable dichotomous right-wrong response by comparing it to a random number between 0 and 1. A response is coded as correct when its associated probability is greater than the random number and incorrect when it is less. This model and procedure is used in the DATAGEN program by Hambleton and Rovinelli (1973) to generate item responses of a single group to an unbiased test.

Item bias was introduced into (1) by making the a_g and b_g item parameters group dependent. That is,

$$P(u_g=1|\theta_{i,j}) = c_g+(1-c_g)\ 1+\exp(-1.7a_{gj}(\theta_i-b_{gj}))^{-1} \qquad (2)$$

where: j indexes group membership

When $a_{g1} \neq a_{g2}$ and/or $b_{g1} \neq b_{g2}$, the probability of a correct response as defined by (2), is also group dependent. Examinees of the same ability, but from different groups have different probabilities of a correct response, and therefore, the item is considered to be biased. The varying amounts and nature of the bias generated for use in the study was manipulated by altering the standard deviation of Δ_{ag} and the standard deviation of Δ_{bg} where $\Delta a_g = a_{g1} - a_{g2}$ and $\Delta b_g = b_{g1} - b_{g2}$.

In this study, data were generated for one hundred twelve different sets of test conditions. Seven test lengths (20, 30, 40,...80 items), four different amounts of bias in discrimination ($\sigma\Delta a_g$ = .0, .4, .8, 1.2) and four different amounts of bias in difficulty ($\sigma\Delta b_g$ = .0, .5, 1.0, 1.5) were employed. The condition of $\sigma\Delta a_g$ - .0 indicated the absence of bias in discrimination and presence solely in item difficulty, while the condition of $\sigma\Delta b_g$ = .0, indicated the absence of bias in difficulty and the presence of bias soley in item discrimination.

Examples of three biased items are shown in Figure 3.1. These figures depict the regression of the probabilities of a correct response on ability for examinees from two different groups, i.e., they illustrate item characteristic curves produced by equation (2). In the first figure, bias was introduced solely into the difficulty parameter ($a_{g1} = a_{g2}$, $b_{g1} \neq b_{g12}$); the second, solely into the discrimination parameter ($a_{g1} \neq a_{g2}$, $b_{g1} = b_{g2}$; and in the third, into both the discrimination and difficulty parameters ($a_{g1} \neq a_{g2}$, $b_{g1} \neq b_{g2}$). For all three of these items, examinees of the

same ability but from different groups have different probabilities of a correct response. The nature and amount of the difference is a function of the dissimilarity of the a_{gj} and b_{gj} parameters. In an unbiased test $a_{g1}=a_{g2}$ and $b_{g1}=b_{g2}$ and consequently, the curves would be identical.

Insert Figure 3.1 About Here

The model defined by (2) was applied first by drawing seed values for the a_g, b_g and c_g parameters in equation (1) from normal distributions N(1.0, .50), N(.0, 1.0) and N(.2, .05), respectively. Second, values for a_g and b_g were drawn from normal distributions with mean values equal to zero. The item parameters for equation (2) were then defined as:

$$a_{gj} = a_g + G_j \, \Delta a_g/2$$

$$\text{and} \quad b_{gj} = b_g + G_j \, \Delta b_g/2 \qquad (3)$$

$$\text{where } G_j = 1 \text{ for } j=1$$
$$G_j = -1 \text{ for } j=2$$

For each general item, the amount of bias was calculated as the area between the item characteristic curves, approximated by:

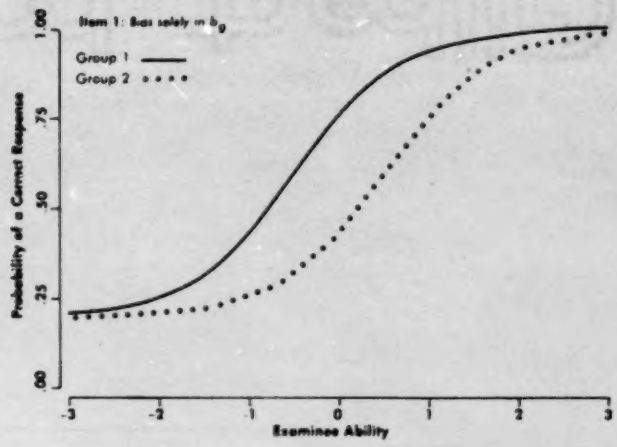
$$\Psi_g = \sum_{-4.000}^{4.000} P(u_g=1|\theta_{i1}) - P(u_g=1|\theta_{i2}) \; \Delta\theta \qquad (4)$$

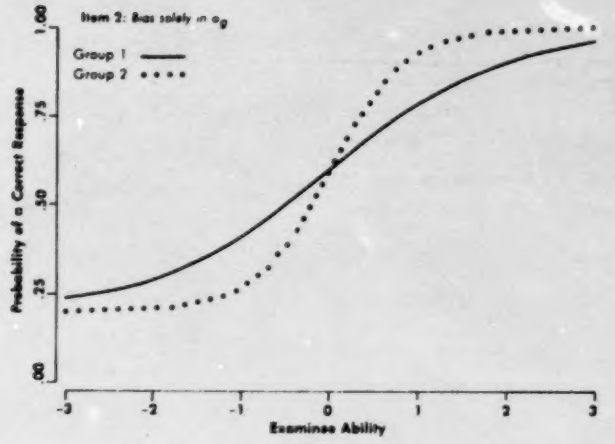
The reader should note that this bias value is independent of the distributions of examinees ability in the two generated groups.

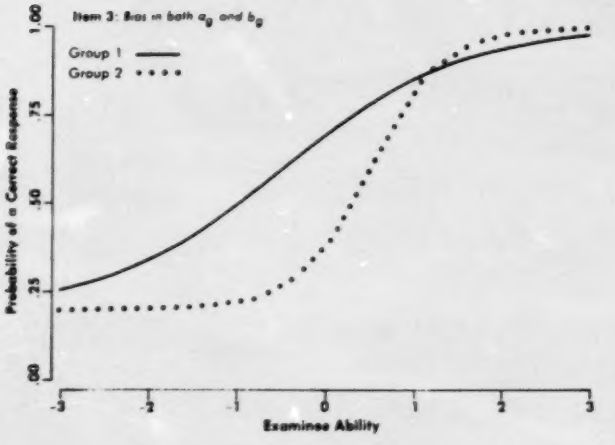
The final step of the process involved drawing two groups of 1200 examinees with different mean levels of ability from normal distributions, N(-.5, 1.0) and N(.5, 1.0). One standard deviation difference in mean level of performance was selected for its correspondence to a value frequently encountered in realistic data of this nature. For each item, the item parameters, the ability of a simulated examinee and the index of group membership were entered into equation (2). The probability of a correct response to the item was thus defined and then transformed to an observable

Figure 3.1:

Hypothetical Item Characteristic Curves for Three Items and Two Groups







right-wrong response. This procedure was repeated under each of the 20, 30, 40,...80 item conditions and for each of the 2400 simulated examinees. The amount of bias generated for each item was then correlated with the amount of bias detected by each of the selected techniques.

Application of the Approaches

The current investigation utilized the transformed item difficulties approach, two applications of item characteristic curve (ICC) theory, and two chi-square approaches to obtain measures of the degree of bias for each item within the generated item pools. The procedures used for the application of these are approaches are as follows:

Transformed Item Difficulties (TID).[1] Two sets of item p-values were computed, one for each population group pair, and transformed to within group normal deviates. From the bvivariate scatterpllot of the sets of transformed p-values, the absolute values of the magnitudes of the item residuals, i.e., the item-45 degree line distances, were computed and used to indicate the relative amount of item bias (see Angoff, 1972).

ICC Theory with Three Parameters (ICC-3). Two sets of ICC parameters, as defined by Birnbaum's three parameter logistic model, were estimated for each of the items by separately applying Urry's (1975) iterative minimum chi-square procedure to the item responses of the two population groups. The parameter value estimates were then equated and the area between estimated equated ICC's served to indicate the extent of item aberrance (see Rudner 1977a).

1. Subroutines for applying the TID, ICC-3 and CHI-5 techniques and for generating biased test items are appended to this report.

ICC Theory with One Parameter (ICC-1). Two sets of item easiness parameters, as defined by Rasch's one parameter logistic model, were estimated for each of the items by separately applying Wright and Mead's (1975) BICAL program to the item responses of the two culture groups. The mean square fit of items to the Rasch model was then computed for each item in the two population groupings. The difference in residuals was used to indicate the relative amounts of item bias (see Durovic, 1975).

Chi-square with Five Intervals (CHI-5). Each item was tested individually for bias using Scheuneman's modified chi-square technique with two culture groups and five total score intervals. With this approach, the expected values for each cell (E_{hj}) were obtained by multiplying the proportion of examinees who respond correctly to the item within a total score interval, by the total number of examinees within the cell that is,

$$E_{hj} = \frac{O_{\cdot j}}{N_{\cdot j}} N_{hj}$$

where: $O_{\cdot j}$ is the number of examinees in total score interval j responding correctly

$N_{\cdot j}$ is the total number of examinees in interval j, and

N_{hj} is the total number of examinees in Group h and score interval j.

As with a conventional chi-square, observed cell values are simply the number of examinees within the cell responding correctly to the item. For each item, the magnitude of aberrance was indicated by the value of the resultant X^2 divided by its associated degrees of freedom, that is a mean square (see Scheuneman, in press, a).

Chi-square with multiple intervals (CHI-N). Each item was tested individually for bias using a modification of the chi-square technique described above. The number of total score intervals was increased to the number of possible total score values, minus the number of cells with expected values less than 5.

CHAPTER IV

RESULTS

Correlations between the amount of bias generated and the amount of bias detected by each technique as a function of test length, are found in Table 1. Each correlation is composed of the items generated under the 16 conditions of $\sigma\Delta a_g$ and $\sigma\Delta b_g$ for each test length. All techniques showed a slight general increase in the average correlation with an increase in test length, except CHI-N. The sharpest increase occurred between twenty and forty items for the ICC-3 technique. In general, however, none of the techniques appeared particularly affected by test length. Table 4.1 also reports the correlations of detected and generated amounts of bias for each technique computed over all 5600 generated test items. The ICC-3, CHI-5 and TID techniques were found to be the most accurate with correlations of .81, .73 and .68 respectively, the ICC-1 technique was the least accurate with a correlation of .56.

Insert Table 4.1 About Here

The correlations between detected and generated amounts of bias by defined level of $\sigma\Delta b_g$ are reported in Table 4.2 The correlations for ICC-3 and TID tended to rise with increases in $\sigma\Delta b_g$ while the ICC-1 correlations exhibited a steady decline. The CHI-5 and CHI-N techniques showed a steady decline in correlations until the extreme condition of $\sigma\Delta b_g = 1.5$.

Insert Table 4.2 About Here

Table 4.1

Correlations of detected and generated amounts of item bias by technique under all 16 conditions of $\sigma\Delta\ a_g$ and $\sigma\Delta\ b_g$ for different test lengths

Test Length	Generated Bias Mean	s.d.	TECHNIQUE TID	CHI-5	CHI-N	ICC-3*	ICC-1
20	.67	.59	.65	.65	.63	.67	.46
30	.67	.62	.70	.74	.62	.81	.49
40	.69	.65	.67	.70	.62	.79	.55
50	.69	.65	.68	.76	.66	.83	.62
60	.69	.61	.66	.72	.68	.81	.54
70	.68	.63	.76	.80	.73	.83	.58
80	.67	.62	.69	.74	.57	.80	.59
All Items	.67	.61	.68	.73	.64	.81	.56

*unparameterized items excluded

Table 4.2

Correlations of detected and generated amounts of item bias by technique for different levels of $\sigma\Delta b_g$ (1400 items)

$\sigma\Delta b_g$	Generated Bias Mean	s.d.	TECHNIQUE TID	CHI-5	CHI-N	ICC-3*	ICC-1
.0	.34	.50	.42	.81	.65	.70	.67
.5	.53	.47	.39	.66	.58	.67	.56
1.0	.77	.52	.61	.66	.58	.73	.53
1.5	1.02	.68	.78	.79	.69	.81	.51

* unparameterized items excluded

The zero condition of $\sigma\Delta b_g$ is the case of bias solely attributable to group differences in item discrimination. Under this condition, the CHI-5 and the ICC-1 techniques were most sensitive to bias with respective correlations of .81 and .70. The lowest correlation was found with the TID.

The correlations between detected and generated amounts of bias by level of $\sigma\Delta a_g$ are reported in Table 4.3 The correlations for the CHI-5, TID, and CHI-N techniques steadily decreased with an increase in a_g whereas the ICC-1 and ICC-3 correlations remained relatively stable. In the condition of zero $\sigma\Delta a_g$, where bias was attributable solely to group differences in item difficulty, the TID and the CHI-5 techniques were the most sensitive with correlations of .87 and .86, respectively. The CHI-5, TID, and CHI-N techniques appeared most influenced by $\sigma\Delta a_g$ differences while the ICC-1 and ICC-3 techniques were less influenced.

Insert Table 4.3 About Here

Inspection of Table 4.4 which reports the intercorrelations among the five investigated techniques reveals that the highest agreement was found between the CHI-5 and TID and between the CHI-5 and CHI-N techniques. The intercorrelations of the ICC-1 and other techniques were consistently the lowest.

Insert Table 4.4 About Here

Table 4.7

Correlation of detected and generated amounts of item bias by technique for different levels of $\sigma\Delta a_g$ (1400 items)

$\sigma\Delta a_g$	Generated Bias Mean	s.d.	TECHNIQUE TID	CHI-5	CHI-N	ICC-3*	ICC-1
.0	.45	.54	.87	.86	.77	.82	.49
.4	.57	.51	.79	.81	.74	.82	.49
.8	.75	.60	.62	.69	.56	.79	.58
1.2	.88	.66	.56	.67	.57	.80	.59

* unparameterized items excluded

Table 4.4

Intercorrelations of detected amounts of item bias (all 5600 items)

	TID	CHI-5	CHI-N	ICC-3*	ICC-1
TID	1.00	.85	.71	.69	.36
CHI-5		1.00	.87	.74	.35
CHI-N			1.00	.66	.32
ICC-3*				1.00	.42
ICC-1					1.00

* unparameterized items excluded

CHAPTER V

DISCUSSION

Since the Rasch model does not include an item discrimination parameter, a biased item detection technique based on the model might be expected to be sensitive mostly to bias in item difficulty. Contrary to this expectation, the ICC-1 approach as advocated by Durovic (1975) was found to be most sensitive to bias in item discrimination. This approach compares goodness- of-fit statistics which in this study were consistently small, indicating good model-data fit for both groups. The best fits were obtained whenever the items of the test were equally discriminating and poorest whenever they were not. Thus, the observed slight differences in item goodness-of-fit statistics were maximized with differences in item discrimination and minimized with differences in item difficulty. Wright, Mead and Draba's (1976) contention that Durovic did not fully exploit the potential of the Rasch model for dealing with the item bias problem was obviously supported.

The three parameter item characteristic curve theory technique used the identical model for detecting item bias as was used in the item generation, and therefore, near perfect correlations of generated and detected amounts of bias could have been expected. The less than perfect correlations are attributable to error in estimation of ability and item parameters. This error was greatest with 20 item tests. Additionally, the correlations were attenuated by the exclusion of approximately seven percent of the items. These were items which could not be parameterized due to extreme a_g or b_g values for one group and tended to have the largest amounts of item bias. Even with this error and attenuation,

the amount of bias detected by this technique (ICC-3) correlated highest among the techniques whenever bias was introduced into both the item difficulty and item discrimination parameters. The potential added cost of item characteristic curve theory item parameterization appears to be offset by the increased accuracy of the techniques.

As pointed out by Scheuneman (in press, b) and empirically supported by Ironson (1978), Rudner and Convey (1978) and this study, the CHI-5 approach is similar to the ICC-3 technique. In the chi-square approach, ability is defined by total observed scores and placed into discrete intervals as opposed to inferred from the item resposes and placed on a continuum. With five total score intervals, the chi-square technique was found to be as effective as the three parameter item characteristic curve theory technique, under most of the investigated conditions.

The increase in the number of intervals in the chi-square technique provided a refinement in the estimate of examinee ability and, hence, theoretically, a refinement in the accuracy of the procedure. In practice however, rather than increasing the accuracy of the technique, there was an overall decrease in accuracy when more total observed score intervals were employed. This appears to be an artifact of using different interval definitions for each item. As pointed out by Rudner and Convey (1978), the contribution of each cell to the X^2 value will no longer be systematic when different interval definitions are employed.

In an excellent critique of item bias detection techniques, Hunter (1975) pointed out that Angoff's transformed item difficulties technique was spuriously sensitive to examinee guessing and differential distributions of examinee ability, yet also insensitive to differences in item discrimination. Hunter's contentions were empirically supported in this investigation, however, the severity of these problems was not found to

be critical. Compared to the other approaches, the overall amount of bias detected by the TID technique correleated relatively well with the generated amounts of item bias - .67, as compared to .72 for CHI-5 and .79 for ICC-3. Since this technique requires only easily obtained within group item p-values and can be applied with the simplest hand calculator, it is definitely worth consideration. At the same time it should be noted that the transformed item difficulties technique was one of the investigated approaches least sensitive to differences in item discrimination.

Since Monte Carlo data with significantly more items and bias was employed in this study, the between-technique correlations of detected amounts of bias which are reported are significantly larger than those found by Ironson (1978) and Rudner and Convey (1978) in empirical investigations using some of the same techniques and actual item response data. With the exception of the ICC-1 application, the intercorrelations of the techniques were all high, although not overwhelming. Different relative amounts of item bias are detected by each of the techniques indicating that they cannot be used interchangeably.

CHAPTER VI

CONCLUSIONS

Item responses for two groups of examinees were simulated for 112 independent tests of various lengths and with various amounts and types of item bias. Five techniques for identifying biased test items were then applied to the item response data. Two of the techniques--item characteristic curve theory with three parameters and chi-square with 5 intervals--produced fairly accurate results under all the investigated conditions. A third technique--transformed item difficulties--was found to be sensitive to bias in item difficulty, but relatively insensitive to bias in item discrimination. Increasing the number of intervals decreased the accuracy of the chi-square technique. The item characteristic curve theory approach with one parameter consistently correlated poorly with the amount of generated bias.

Regardless of test length, amount of bias or nature of bias, the item characteristic curve theory with three parameters, the chi-square technique with 5 intervals is to be recommended for detecting biased test items. As an alternative or when only item p-values are available, the transformed item difficulties technique may be used. However, it must be recognized that this approach will fail to detect bias in item discrimination.

APPENDIX

```
C
C
C
C
C
C
C       THIS APPENDIX CONTAINS SUBROUTINES FOR
C
C
C               BIASED ITEM GENERATION                  (BIGEN )
C
C               BIASED ITEM DETECTION VIA
C
C                       TRANSFORMED ITEM DIFFICULTIES (REGPLT)
C                       CHI-SQUARE WITH N INTERVALS   (SCHI  )
C                       THREE PARAMETER ICC TECHNIQUE (ICBIAS)
C
C               OMITTED ARE PARAMTERIZATION VIA THE ONE AND THREE
C               PARAMETER ICC TECHNIQUES.  THE SUBROUTINES MISSING
C               ARE ADAPTATIONS OF URRY'S ANCILLES AND WRIGHT'S BICAL
C               PROGRAMS.
C
C
C
C       MAIN    VERSION 5.3
C       RUNS AS MAIN CALLER FOR BIASED ITEM PROGRAMS
C
        DIMENSION IDA(2,1200,3),  ACB(80), XFIT(80),
        1 DITEM1(80), DITEM2(80), BTID(80)
        2 ,AREA(80), S5(80), SNI(80)
        open(unit=21,file="biased",DEVICE="DSK",ACCESS="SEQOUT")
C       NUMBER OF SUBJECTS
        NS=1200
C
C       SELECT NUMBER OF ITEMS
C
        DO 30 NI=20,80,10
C
C       IB=5
        DO 30 IB=0,15,5
C       SELECT SD OF BIAS IN DIFFICULTY
C
        BDTOP=IB/10.
C
C       SELECT SD OF BIAS IN DISCRIMINATION
C
        DO 30 IA=0,12,4
        BDITOP=IA/10.
C
C       GENERATE DATA
        CALL BIGEN(NI,NS,BDTOP,BDITOP,IDA,AREA,DITEM1,
        1 DITEM2)
C
C       ANALYZE USING DUROVIC S TECHNIQUE
        CALL BIRAS(NS,NI,IDA,XFIT)
C       TRANSFORMED ITEM DIFFICULTIES
        CALL REGPLT(NI,DITEM1,DITEM2,BTID,R,0)
C       DO LATENT TRAIT THEORY
        CALL ICBIAS(NI,NS,NS,IDA,ACB )
C       CHI-SQUARE WITH 5 INTERVALS
```

```
      CALL SCHI(NI,NS,5,IDA,S5)
C     CHI-SQUARE WITH N INTERVALS
      CALL SCHI(NI,NS,NI,IDA,SNI)
CCC   FORMAT(' ITEM  DIF    DIS SIZE  BIAS TID   X-5   X-N   ICC',/)
      DO 20 I=1,NI
      WRITE(21,10) I,BDTOP,BDITOP, NI, AREA(I),   BTID(I), S5(I), SNI(I)
      1, ACB(I), XFIT(I)
10    FORMAT(I5,2F6.2,I5,6F6.2)
20    CONTINUE
30    CONTINUE
      END
C*****************************************************************
C*****************************************************************
C
C
C
      SUBROUTINE BIGEN(NI,NS,BDTOP, BDITOP,IDA,AREA,DITEM1,
      1 DITEM2)
      DIMENSION AA(80), BB(80), CC(80), DITEM1(NI), DITEM2(NI),
      1   IP(80), F1(81), F2(81), AREA(80), IDA(2,NS,3),
      2 I2(3), BBIAS(80), ABIAS(80)
      NE1=NS
      NE2=NS
      IDIF=1
      DBOT=0
      DTOP=1
      IDISC=1
      DIBOT=1.
      DITOP=.5
      IGU=1
      GBOT=.20
      GTOP=.05
      IS1=123245
      IBDIF=1
      BDBOT=0.0
      IBDIS=1.0
      BDIBOT=0.0
      IFORM1=1
      IFORM2=1
      ABOT1= .50
      ATOP1=1.00
      ABOT2= -.5
      ATOP2=1.0
      CALL RANSET(IS1,IS2,IS3,IS4,IS5,IS6,IS7)
      EN1 = NE1
      EN2 = NE2
      KK = NI+1
      FN = NI
      DO 10 I=1,NI
      AREA(I)=0.0
      ABIAS(I)=0.0
      BBIAS(I)=0.0
      DITEM1(I) = 0.0
      DITEM2(I) = 0.0
      F1(I) = 0
10    F2(I) = 0
      F1(KK) = 0
      F2(KK) = 0
C
      RD = DTOP - DBOT
```

```
      RG = GTOP - GBOT
      RDIS = DITOP - DIBOT
      RBD = BDTOP - BDBOT
      RBDI = BDITOP - BDIBOT
C
C     GENERATE ITEM PARAMETERS
C
      CALL GENPAM(BB,NI,IDIF,DBOT,DTOP,IS1,-3.,3.)
      CALL GENPAM(AA,NI,IDISC,DIBOT,DITOP,IS2,.5,3.)
      CALL GENPAM(CC,NI,IGU,GBOT,GTOP,IS3,.0,1.)
      CALL GENBIA(BBIAS,BB,NI,IBDIF,BDBOT,BDTOP,IS4,-3.,3.)
      CALL GENBIA(ABIAS,AA,NI,IBDIS,BDIBOT,BDITOP,IS5,.0,3.)
      CALL BIAREA(AA,BB,CC,BBIAS,ABIAS,NI,AREA)
      DO 20 I=1,NI
      A1=AA(I) - ABIAS(I)/2.
      A2=AA(I) + ABIAS(I)/2.
      B1=BB(I) - BBIAS(I)/2.
      B2=BB(I) + BBIAS(I)/2.
20    CONTINUE
C
C     GENERATE ITEM RESPONSES FOR GROUP 1
C
      CON = ATOP1 - ABOT1
      DO 80 I=1,NE1
      IR = 0
      GO TO (30,40) IFORM1
30    CALL GAUSS(ABIL,IS5)
      ABIL= ATOP1*ABIL + ABOT1
      GO TO 50
40    ABIL=RAN(IS5)
      ABIL= CON*ABIL + ABOT1
50    DO 60 J=1,NI
      A1 = AA(J) - ABIAS(J)/2.
      B1 = BB(J) - BBIAS(J)/2.
      TT=RAN(IS7)
      PP=PROBT(A1,B1,CC(J),ABIL)
      IP(J) = 0
      IF (TT.GT.PP )    GO TO 60
      IP(J) = 1
      DITEM1(J) = DITEM1(J) + 1
      IR = IR + 1
60    CONTINUE
      CALL PAK(NI,IP,I2)
      DO 70 J=1,3
70    IDA(1,I,J)=I2(J)
      IR = IR + 1
      F1(IR) = F1(IR) + 1.
80    CONTINUE
C
C     GENERATE ITEM RESPONSES FOR GROUP 2
C
      CON = ATOP2 - ABOT2
      DO 140 I=1,NE2
      IR = 0
      GO TO (90,100) IFORM2
90    CALL GAUSS(ABIL,IS6)
      ABIL = ATOP2*ABIL + ABOT2
      GO TO 110
100   ABIL=RAN(IS6)
      ABIL= CON*ABIL + ABOT2
```

```
110     DO 120 J=1,NI
        A1 = AA(J) + ABIAS(J)/2.
        B1 = BB(J) + BBIAS(J)/2.
        TT=RAN(IS7)
        PP=PROBT(A1,B1,CC(J),ABIL)
        IP(J) = 0
        IF (TT.GT.PP )    GO TO 120
        IP(J) = 1
        DITEM2(J) = DITEM2(J) + 1
        IR = IR + 1
120     CONTINUE
        CALL PAK(NI,IP,I2)
        DO 130 J=1,3
130     IDA(2,I,J)=I2(J)
        IR = IR + 1
        F2(IR) = F2(IR) + 1.
140     CONTINUE
        SUM1 = .0
        SUM2 = .0
        DO 150 J=1,NI
        DITEM1(J) = DITEM1(J)/EN1
        DITEM2(J) = DITEM2(J)/EN2
150     CONTINUE
        RETURN
        END
C*****************************************************************
C
C
        FUNCTION PROBT(A,B,C,TH)
        PROBT= C + (1.-C)/(1. + EXP(-1.7*A*(TH-B) ) )
        RETURN
        END
C
C*****************************************************************
C
        SUBROUTINE GAUSS(V,ISEED)
        A = .0
        DO 10 J=1,12
        XX = RAN(ISEED)
10      A = A + XX
        V = A - 6.
        RETURN
        END
C
C*****************************************************************
C
        SUBROUTINE BIAREA(AA,BB,CC,BBIAS,ABIAS,NI,AREA)
C       COMPUTES AREA BETWEEN TWO ICC S
        DIMENSION AA(NI), BB(NI), CC(NI),BBIAS(NI),ABIAS(NI),AREA(NI)
        DO 10 I=1,NI
        A1 = AA(I) - ABIAS(I)/2.
        A2 = AA(I) + ABIAS(I)/2.
        B1 = BB(I) - BBIAS(I)/2.
        B2 = BB(I) + BBIAS(I)/2.
        C= CC(I)
        DO 10 J=-400,400,5
        TH = J/100.
10      AREA(I) = AREA(I) + ABS(PROBT(A1,B1,C,TH) - PROBT(A2,B2,C,TH) )
        1  * .05
        RETURN
```

```
      END
C
C*****************************************************************
C
      SUBROUTINE GENPAM(RR,NI,IND,BOT,TOP,ISEED,ALOW,AHIGH)
      DIMENSION RR(NI)
      GO TO (10,40) IND
10    DO 30 I = 1,NI
20    CALL GAUSS(RR(I),ISEED)
      RR(I) = TOP*RR(I) + BOT
      IF (RR(I).LT.ALOW .OR. RR(I).GT.AHIGH)    GO TO 20
30    CONTINUE
      RETURN
40    CON = TOP - BOT
      DO 50 I=1,NI
      X = RAN(ISEED)
      RR(I) = CON * X + BOT
50    CONTINUE
      RETURN
      END
C
C*****************************************************************
C
      SUBROUTINE GENBIA(RR,AA,NI,IND,BOT,TOP,ISEED,ALOW,AHIGH)
      DIMENSION RR(NI),AA(NI)
      GO TO (10,40) IND
10    DO 30 I=1,NI
20    CALL GAUSS(RR(I),ISEED)
      RR(I) = TOP*RR(I) + BOT
      A = ABS(RR(I)/2.0 )
      CH1 = AA(I) - A
      CH2 = AA(I) + A
      IF (CH1.LT.ALOW .OR. CH2.GT.AHIGH)    GO TO 20
30    CONTINUE
      RETURN
40    CON = TOP - BOT
      DO 60 I = 1,NI
50    X = RAN(ISEED)
      RR(I) = CON*X + BOT
      A = ABS( RR(I)/2. )
      CH1 = AA(I)-A
      CH2 = AA(I)+A
      IF( CH1.LT.ALOW .OR. CH2.GT.AHIGH)    GO TO 50
60    CONTINUE
      RETURN
      END
C
C*****************************************************************
C
      SUBROUTINE RANSET(IS1,IS2,IS3,IS4,IS5,IS6,IS7)
      IS1 = 1000*RAN(IS1)
      IS2 = 1000*RAN(IS1)
      IS3 = 1000*RAN(IS2)
      IS4 = 1000*RAN(IS3)
      IS5 = 1000*RAN(IS4)
      IS6 = 1000*RAN(IS5)
      IS7 = 1000*RAN(IS6)
      RETURN
      END
C*****************************************
```

```
C*****************************************************************
C
      SUBROUTINE UNPAK(NI,I1,I2)
      DIMENSION I1(NI),I2(3)
      IV= (NI-1)/27. + 1
      DO 10 I=1,NI
10    I1(I)=0
      IM=27
      DO 20 I=IV,1,-1
      IT= NI-27*(I-1)
      IM= MIN0(IT,27)
      DO 20 J=IM,1,-1
      DEM = 2.**(J-1)
      IN= (I-1)*27. + J
      I1(IN)= I2(I)/DEM
      I2(I)=I2(I) - I1(IN)*DEM
20    CONTINUE
      RETURN
      END
C*********************************
C*****************************************************************
C
      SUBROUTINE PAK(NI,I1,I2)
      DIMENSION I1(NI),I2(3)
      DO 10 I=1,3
10    I2(I)=0
      K=0
      J=0
      DO 20 I=1,NI
      J=J+1
      IF(J.EQ.28)  J=1
      IF(J.EQ.1)  K=K+1
      I2(K)=I2(K)+I1(I)*( 2.**(J-1) )
20    CONTINUE
      RETURN
      END
C*****************************************************************
C*****************************************************************
C
C
C
      SUBROUTINE BIRAS(NS,NI,IDA,XFIT)
      DIMENSION IDA(2,NS,3), XFIT(NI),XFIT2(80)
      CALL BICAL(NS,NI,IDA,1,XFIT)
      CALL BICAL(NS,NI,IDA,2,XFIT2)
      DO 10 I=1,NI
      XFIT(I)= ABS( XFIT(I)-XFIT2(I))
10    CONTINUE
      RETURN
      END
C*****************************************************************
C*****************************************************************
C
C
C
      SUBROUTINE REGPLT(NI,VP1,VP2,BTID,R,IGD)
C     INPUT
C             NI  -  NUMBER OF ITEMS
C             VP1(NI) -  P-VALUES GROUP1
C             VP2(NI) -  P-VALUES GROUP 2
```

```
C       OUTPUT
C               BTID(NI) - BIAS IN THE ITEM
        DIMENSION VP1(NI),VP2(NI), BTID(NI)
10      I=0
C ***       ZEROS STORAGE VECTORS
        NN=0
        AN=0.0
        SX=0
        SY = 0
        SX2=0
        SY2=0
        SXY=0
        DO 20 I=1,NI
        IF(IGD.EQ.1 .AND. VP2(I).EQ. .0)  GO TO 20
        AN=AN+1
        X=VP1(I)
        Y=VP2(I)
        SX=SX + X
        SY = SY + Y
        SX2 = SX2 + X**2
        SY2 = SY2 + Y**2
        SXY = SXY + X*Y
20      CONTINUE
C ***       COMPUTES CORRELATIAN
30      R =(SXY - (SX*SY)/AN) /
        1 (SQRT( (SX2 -(SX**2)/AN)*(SY2-(SY**2)/AN) ) )
C ***       COMPUTES MEANS
        XM=SX/AN
        YM=SY/AN
C ***       COMPUTES STANDARD DEVIATIANS
        XSD = SQRT(SX2/AN - XM**2)
        YSD = SQRT(SY2/AN - YM**2)
C ***       COMPUTES Z-SCORES, ITEM-REGRESSIAN LINE DISTANCES AND
C ***       PLOTS LOCATIANS
        DO 40 I=1,NI
        X=VP1(I)
        Y=VP2(I)
        ZX= (X-XM)/XSD
        ZY = (Y-YM)/YSD
        D = (ZY - ZX) / 1.41421356
        BTID(I)=D
40      CONTINUE
        RETURN
        END
C*************************************************************************
C*************************************************************************
C
C
C
C
        SUBROUTINE ICBIAS(KITEMS, NS1,NS2, IDA,BIAS)
C       ICBIAS.SUB      VERSION 1.0
C       DOES ITEM BIAS DETECTION VIA LATENT TRAIT THEORY
C       TECHNIQUE
C
C       SUBROUTINES REQUIRED ARE: PARAM, AREA, ICEQU
C       FUNCTION REQUIRED: PROBT
C       KITEMS  - NUMBER OF ITEMS
C       NS1     - NUMBER OF SUBJECTS GROUP 1
C       NS2     - NUMBER OF SUBJECTS GROUP 2
```

```
C       IDA     - PACKED DATA FOR THE TWO GROUPS
C       BIAS    - RETURNED AMOUNT OF BIAS FOR EACH ITEM
C
        DIMENSION IDA(2,1200,3), I2(3), ABC1(3,80),ABC2(3,80)
        DIMENSION BIAS (80)
        CALL PARAM(KITEMS, NS1, 1, IDA, ABC1)
        CALL PARAM(KITEMS, NS2, 2, IDA, ABC2)
C       ABC1 AND ABC2 CONTAIN THE ABC PARAMETER VALUES FOR
C       ALL ITEMS FOR GROUPS 1 AND 2.  NOW TO EQUATE THE
C       ICC PARAMETER VALUES
        CALL ICEQU(ABC1,ABC2,KITEMS)
C       NOW TO COMPUTE THE AREA BETWEEN THE EQUATED ICC'S
        CALL AREA(ABC1,ABC2,KITEMS,BIAS)
        RETURN
        END
C*************************************************************************
        SUBROUTINE AREA(ABC1,ABC2,NI,BIAS)
        DIMENSION ABC1(3,NI), ABC2(3,NI), BIAS(NI)
        DO 20 I=1,NI
        BIAS(I)=0.0
        A1=ABC1(1,I)
        B1=ABC1(2,I)
        C1=ABC1(3,I)
        A2=ABC2(1,I)
        B2=ABC2(2,I)
        C2=ABC1(3,I)
        IF( A1.EQ. .0  .AND. B1.EQ. .0)  GO TO 20
        DO 10 J=-400,400,5
        TH=J/100.
        BIAS(I)=BIAS(I) + (ABS (PROBT(A1,B1,C1,TH)-PROBT(A2,B2,C2,TH) ))
        1 * .05
10      CONTINUE
20      CONTINUE
        RETURN
        END
C*************************************************************************
        SUBROUTINE ICEQU(ABC1,ABC2,NI)
C       EQUATES ICC'S
        DIMENSION ABC1(3,NI),ABC2(3,NI)
C       COMPUTE THE STANDARD DEVIATIONS OF B1, B2,A1,A2
C       AND THE MEANS OF B1 AND B2
C       ITEMS WITH A LT .8 OR ABS(B) GT 2.0 WILL BE DROPPED
        ANI1= 0.0
        ANI2=0.0
        XB1=0.0
        XB2=0.0
        SSQB1=0.0
        SSQB2=0.0
        DO 30 I=1,NI
        IF (ABC1(1,I).EQ. .0 .AND. ABC1(2,I).EQ..0)  GO TO 10
        GO TO 20
10      ABC2(1,I)=0.0
        ABC2(2,I)=0.0
        ABC2(3,I)=0.0
20      IF( ABC1(1,I).LT..5)  GO TO 30
        ABB1=ABS( ABC1(2,I) )
        IF (ABB1.GT. 2.5)  GO TO 30
        XB1=XB1+ABC1(2,I)
        SSQB1=SSQB1+ABC1(2,I)**2
        ANI1=ANI1+1
```

```
30      CONTINUE
        DO 60 I=1,NI
        IF ( ABC2(1,I).EQ.0 .AND. ABC2(2,I).EQ.0)  GO TO 40
        GO TO 50
40      ABC1(1,I)=0.0
        ABC1(2,I)=0.0
        ABC1(3,I)=0.0
50      IF( ABC2(1,I).LT..5)  GO TO 60
        ABB1=ABS( ABC2(2,I) )
        IF( ABB1.GT. 2.5)  GO TO 60
        XB2=XB2+ABC2(2,I)
        SSQB2=SSQB2+ABC2(2,I)**2
        ANI2=ANI2+1
60      CONTINUE
        XB1=XB1/ANI1
        XB2=XB2/ANI2
        SDB1= ( SSQB1/ANI1 - XB1**2)
        SDB2= ( SSQB2/ANI2 - XB2**2)
        SDB1=SQRT(SDB1)
        SDB2=SQRT(SDB2)
C
        SDB1= SDB1/SDB2
        DO 70 I=1,NI
        IF( ABC2(1,I).EQ..0 .AND. ABC2(2,I).EQ..0)  GO TO 70
        ABC2(2,I) = ABC2(2,I)*SDB1 + (XB1 - SDB1*XB2)
        ABC2(1,I)= ABC2(1,I)/SDB1
        ABC2(3,I)=ABC1(3,I)
70      CONTINUE
        RETURN
        END
C*******************************************************************
C*******************************************************************
C
C
C
        SUBROUTINE SCHI(NI,NS,NINT,IDA,CHISQ)
C
C       INPUT -
C               NI      - NUMBER OF ITEMS
C               NS      - NUMBER OF SUBJECTS
C               NINT    - NUMBER OF DESIRED INTERVALS
C               IDA(2,NS,3) - PAKED RESPONSE DATA
C
C       OUTPUT -
C               CHISQ(NI)       - BIAS BY ITEM
C
C
        DIMENSION TOTP(2,80),TOTR(2,80),FREQ(2,80),OBS(2,80),
        1    AEXP(2,80), DIS(2,80,80), CH(80), CHISQ(80),
        2 I1(80), INT(200), IDA(2,NS,3), I2(3)
        NITEMS=NI
        IMAX=2 * NINT
        DO 10 I=1,80
        CH(I)=0
        CHISQ(I)=0
        DO 10 IP=1,2
        FREQ(IP,I)=0
        TOTP(IP,I)=0
        TOTR(IP,I)=0
        OBS(IP,I) =0
```

```
      AEXP(IP,I)=0
      DO 10 J=1,80
      DIS(IP,I,J)=0
10    CONTINUE
      AT = NINT
C ***
C ***   USER INPUTS FORMAT INFO ON EACH ITEM RESPONSE FILE
C ***
C ***
C ***   READS DATA FILES DELEVOPS NITEMS X NITEMS MATRIX
C ***   OF TOT SCORE BY ITEM WITH ENTRIES BEING THE NUMBER
C ***   OF EXAMINEES RESPONDING CORRECTLY TO THE ITEM
C ***   AND ARRAYS OF THE TOTAL NUMBER OF PEOPLE AT
C ***   EACH TOTAL SCORE LEVEL.
C ***
      DO 40 II=1,NS
      DO 20 J=1,3
      I2(J)=IDA(1,II,J)
20    CONTINUE
      CALL UNPAK(NI,I1,I2)
      NT= SUM(NI,I1)
      IF(NT.EQ.0 .OR. NT.EQ.NI)  GO TO 40
      DO 30 I = 1,NITEMS
      IF (I1(I).NE.1)    GO TO 30
      DIS(1,NT,I) = DIS(1,NT,I) + 1.0
      TOTR(1,I) = TOTR(1,I) + 1.0
30    CONTINUE
      TOTP(1,NT) = TOTP(1,NT) + 1.0
40    CONTINUE
      DO 70 II=1,NS
      DO 50 J=1,3
      I2(J)=IDA(2,II,J)
50    CONTINUE
      CALL UNPAK(NI,I1,I2)
      NT=SUM(NI,I1)
      DO 60 I = 1,NITEMS
      IF (I1(I).NE.1)    GO TO 60
      DIS(2,NT,I) = DIS(2,NT,I) + 1.0
      TOTR(2,I) = TOTR(2,I) + 1.0
60    CONTINUE
      TOTP(2,NT) = TOTP(2,NT) + 1.0
70    CONTINUE
C ***
C ***   ***  DEFINES INTERVALS FOR EACH ITEM BASED ON THE
C ***        AVERAGED ITEM SCORE DISTRIBUTIONS
C ***
      INT(IMAX)=NITEMS
      INT(1)=0
      DO 190 I=1,NITEMS
      NAMIN=0
      AMIN=0
      LINT=NINT
80    LINT=LINT-AMIN
      AT=LINT
      ITMP = LINT - 1
      ITMP2 = 2*ITMP
      DSUM = 0.
      DO 90 K=1,LINT
      DO 90 KG=1,2
      AEXP(KG,K) = 0.0
```

```
          OBS(KG,K)  = 0.0
          FREQ(KG,K) = 0.0
 90       CONTINUE
          DO 100 KK = 1,ITMP
          CH(KK) = KK / AT
 100      CONTINUE
          DO 130 J = 1,NITEMS
          DSUM = DSUM + ( (DIS(1,J,I)/TOTR(1,I) ) +
          1  (DIS(2,J,I)/TOTR(2,I) ) )
          2              / 2.0
          DO 110 KK = 1,ITMP
          KL = KK *2
          IF (DSUM .LE. CH(KK) )    INT(KL) = J
 110      CONTINUE
          DO 120 KL = 2,ITMP2,2
          KK = KL + 1
          INT(KK) = INT(KL) + 1
 120      CONTINUE
 130      CONTINUE
 C ***    COMPUTES FREQUENCY AND OBS VALUES
 C ***
          DO 160 J = 1,NITEMS
          DO 150 K=1,LINT
          KK = 2.0 * K
          KL = KK -1
          IF (J.LT.INT(KL).OR. J.GT. INT(KK) )    GO TO 150
          DO 140 IG=1,2
          FREQ (IG,K) = FREQ(IG,K) + TOTP(IG,J)
          OBS(IG,K) = OBS(IG,K) + DIS(IG,J,I)
 140      CONTINUE
 150      CONTINUE
 160      CONTINUE
          CHISQ(I)= 0.0
 C ***
 C *** **    COMPUTES AEXPECTED VALUES
 C ***
          DO 170 K=1,LINT
          DO 170 KG=1,2
          AEXP(KG,K)= FREQ(KG,K) * (OBS(1,K) + OBS(2,K) ) /
          1                      ( FREQ(1,K) + FREQ(2,K) )
          IF(AEXP(KG,K) .GE. 5.0)  GO TO 170
          AMIN=1
          NAMIN=NAMIN+1
          GO TO 80
 170      CONTINUE
 C ***
 C *** **    COMPUTES MODIFIED CHI SQUARE
 C ***
          DO 180 K=1,LINT
          DO 180 KG=1,2
          CHISQ(I)=CHISQ(I) + ((AEXP(KG,K) - OBS(KG,K))**2)/
          1                ( AEXP(KG,K) )
 180      CONTINUE
          CHISQ(I)=CHISQ(I)/ (AT-1.)
 190      CONTINUE
          RETURN
          END
 C*****************************************************************
          FUNCTION SUM(NI,I1)
 C        SUMS THE NUMBER OF CORRECT RESPONSES IN I1
```

```
      DIMENSION I1(NI)                                              51
      SUM=0
      DO 10 I=1,NI
      IF(I1(I).EQ.1)  SUM=SUM+1
10    CONTINUE
      RETURN
      END
```

Authors

Rudner, Lawrence M.

Address: Gallaudet College
Washington, D.C. 20002

Title: Senior Research Specialist

Degrees: B.S., Syracuse University; M.S., Western Ct. State College,
Ph.D., Catholic University of America.

Specialization: Psychometric Theory, Computer Applications.

Getson, Pamela R.

Address: Gallaudet College
Washington, D.C. 20002

Title: Senior Research Specialist

Degrees: B.S., University of Illinois; M. Ed., University of Virginia;
Ph.D., University of Virginia.

Specialization: Educational Measurement, Research Methodology.

Knight, David L.

Address: Gallaudet College
Washington, D.C. 20002

Title: Senior Research Specialist

Degrees: B.A., Dakota Western University; M.A., University of Miami;
Ph.D. candidate Catholic University of America.

Specialization: Educational Research, Deafness Research.

References

Anastasi, A. Psychological Testing (4th ed.). New York: MacMillan, 1976.

Angoff, W. H. A technique for the investigation of cultural differences. Paper presented at the annual meeting of the American Psychological Association, Honolulu, May 1972.

Angoff, W. H., & Ford, S. F Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement. 1973, 10, 95-105.

Angoff, W. H., & Modu, C. C. Equating the scales of the Prueba de Aptitud Academica and the Scholastic Aptitude Test. New York: College Entrance Examination Board, 1973.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novik, Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley, 1968, Chapters 17-20.

Cardall, C., & Coffman, W. R. A method for comparing performance of different groups on the items in a test. (RM 64-61) Princeton: Educational Testing Service, 1964.

Cleary, T. A. Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 1968, 5, 115-124.

Cleary, T. A., & Hilton, T. L. An investigation into item bias. Educational and Psychological Measurement, 1968, 8, 61-75.

Cole, N. S. Bias in selection. Journal of Educational Measurement, 1973, 10, 237-255.

Darlington, R. B. Another look at "culture fairness." Journal of Educational Measurement, 1971, 8, 71-82.

Durovic, J. Definitions of test bias: A taxonomy and an illustration of an alternative model. Unpublished doctoral dissertation, State University of New York at Albany, 1975.

Echternacht, G. A quick method for determining test bias. Educational and Psychological Measurement. 1974, 34, 271-280.

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., & Tyler, R. W. Intelligence and Cultural Differences. Chicago: University of Chicago Press, 1951.

Einhorn, H. J., & Bass, A. R. Methodological considerations relevant to discrimination in employment testing. Psychological Bulletin, 1971, 75, 261-269.

Fishbein, R. L. An investigation of the fairness of the items in a test battery. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, D.C. April, 1975.

Green, D. R. Reducing bias in achievement tests. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April, 1976.

Green, D. R., & Draper, J. F. Exploratory Studies of Bias in Achievement Tests. Monterey: CTB/McGraw-Hill, 1972.

Gross, A. L., & Su, W. Defining a "fair" or "unbiased" selection model: A question of utilities. Journal of Applied Psychology, 1975, 60, 345-351.

Hambleton, R. K. & Rovinelli, R. A. Fortran IV program for generating item response data from logistic trest models. Behavioral Science, 1973, 18, 74.

Hicks, M. M., Donlon, T. F., & Wallmark, M. M. Sex differences in item responses on the Graduate Record Examination. Paper presented at the Annual Meeting of the National Council of Measurement in Education, San Francisco, April 1976.

Hoepfner, R., & Strickland, G. P. Investigating Test Bias. Los Angeles: Center for the Study of Evaluation, University of California, 1972.

Hunter, J. E. A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items. Paper presented at the NIE Invitational Conference on Test Bias, Annapolis, 1975.

Ironson, G. A comparative analysis of several methods of assessing item bias. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, 1978.

Jensen, A. P. An examination of culture bias in the Wonderlic Personnel Test. Arlington, VA: ERIC Clearinghouse, 1973 (ERIC Document Reproduction Service ED 086 726).

Ku, H. H., & Kullback, S. Loglinear models in contingency table analysis. The American Statistician, 1974, 28, 115-122.

Lord, F. M. A study of item bias using item characteristic curve theory. Proceedings of the Third Congress of Cross Cultured Psychology, Tilburg, Holland, 1977.

Lord, F.M., & Novick, M. R. Statistical Theories of Mental Test Scores (2nd ed.). Reading, MA: Addison-Wesley, 1974.

Maw, C. Item bias and information in item responses. Paper presented at the Psychometric Society Meeting, June 1977.

Merz, W. R. Estimating bias in test items utilizing principal components analysis and the general linear solution. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 1976.

Merz, W. R. Test Fairness and Test Bias: A Review of Procedures in M. Wargop and D. R. Green, Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation, Monterey: McGraw Hill, 1978.

Ozenne, D. G., Van Gelder, N. C., & Cohen, A. J. Energizing School Aid Act (ESAT) National Evaluation, Achievement Test Restandardization. Santa Monica, CA: Systems Development Corporation, 1974.

Pine, S. M. Applications of item characteristic curve theory to the problems of test bias. In D. J. Weiss (Ed.) Applications of Computerized Adaptive Testing (RR 77-1). Minneapolis: University of Minnesota Psychometric Methods Program, March 1977.

Plake, B. S., & Hoover, H. D. An analytical method of identifying biased test items. Paper presented at the Annual Meeting of the American Educational Research Association. New York, April 1977.

Rudner, L. M. An approach to biased item identification using latent trait measurement theory. Paper presented at the Annual Meeting of the American Educational Research Association, New York, April 1977a.

Rudner, L. M. Efforts toward the development of unbiased selection and assessment instruments. Paper presented at the Third International Symposium on Educational Testing, University of Leyden, The Netherlands, 1977b.

Rudner, L. M. Using standard tests with the hearing impaired: The problem of item bias. Volta Review, 1978, 80, 31-40.

Rudner, L. M. & Convey, J. J. An evaluation of select approaches for biased item identification. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, 1978.

Scheuneman, J. A new method of assessing bias in test items. Journal of Educational Measurement, 1979, in press.

Strassberg-Rosenberg, B., & Donlon, T. F. Context influences on sex differences in performance and aptitude tests. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, D.C., 1975.

Thorndike, R. L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8, 63-70.

Thurstone, L. L. A method of scaling psychological and educational tests. Journal of Education and Psychology, 1925, 16, 433-451.

Tucker, L. R. An interbattery method of factor analysis. Psychometrika, 1958, 23, 111-136.

Urry, V. W. Ancillary estimators for the parameters of mental test models. Paper presented at the American Psychological Association Convention, Chicago, August 1975.

Veale, J. R., & Foreman, D. I. Cultural validity of items and tests: A new approach. Score Technical Report, Iowa City, Iowa: Westinghouse Learning Corporation/Measurement Research Center, 1975.

Veale, J. R., & Foreman, D. I. Cultural variation in criterion-referenced tests: A "global" item analysis. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 1976.

Williams, R. L., Mosby, D., & Hinson, V. Critical issues in achievement testing children from diverse ethnic backgrounds. In M. Wargo and D. R. Green, Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation, Monterey: McGraw Hill, 1978.

Wright, B., Mead, R., & Draba, R. Detecting and correcting test item bias with a logistic response model. Chicago: Department of Education, University of Chicago, October, 1976, RM-22.